Universidade do Minho
Escola de Engenharia
Mestrado em Engenharia Informática

Análise Inteligente de Sistemas de Big Data

Ano lectivo 2023/2024

The impact of the purchasing power in mental health and productivity in different countries

Grupo 1

Ana Rita Santos Poças, PG53645

Nelson José Marques Martins Almeida, PG52697

6 de Junho de 2024

Index

Image index

1 Introduction

This report aims to provide a comprehensive overview of the methodology and analysis conducted as part of a practical assignment for the course on Intelligent Analysis in Big Data Systems.

The focus of the assignment was to investigate the relationship between purchasing power, mental health, and productivity across various countries. The initial phase involved meticulously selecting datasets that contained relevant information for the analysis. Subsequently, a rigorous data preparation process was undertaken to ensure the seamless integration of the datasets. The datasets were then merged into a cohesive database to facilitate in-depth analysis.

Finally, the combined data was visualized using advanced visualization techniques to extract meaningful insights. This process allowed for a nuanced exploration of the impact of purchasing power on mental health and productivity, shedding light on potential correlations and trends within the data.

2 State of the Art

Over the past few years, there has been a growing recognition of the intricate interplay between economic factors, mental health outcomes, and productivity levels on a global scale. As nations pursue prosperity, understanding the link between purchasing power, well-bbeing, and productivity is crucial.

Our study aims to explore the complex relationship between purchasing power and its effects on mental health and productivity dynamics in various countries. By examining how fluctuations in purchasing power affect mental health and productivity, we aim to highlight key aspects of socio-economic well-being and inform discussions at the intersection of economics and public health.

To demonstrate the importance of our research, we reviewed existing literature.

The article [1] investigated the economic burden of mental illnesses on purchasing power among Nigerian outpatients. Researchers interviewed 284 outpatients at a psychiatric hospital to understand the financial burden of mental illness. The interviewed focused on both direct costs(e.g, hospital admissions) and indirect costs(e.g, lost productivity) associated with various mental health diagnoses.

The study found mental illness creates a significant financial burden for Nigerians, especially those struggling financially. Treatment costs and lost productivity make basic necessities difficult to afford. Wealthier participants fared better, highlighting unequal access to care. The financial burden worsens health outcomes. This study suggests a longer-term analysis and standardized methods for future research, and advocates for mandatory health insurance, particularly for mental health, to ease the financial strain.

The study [5] examined mental health interventions for working-age adults (18-65) and their impact on work productivity. Excluding volunteers and caregivers, it looked for links between interventions and measures like absenteeism, presenteeism, and job loss.

While the study found a clear connection between poor mental and lower productivity, it highlighted limitations: lack of high quality, long term studies and difficulty isolating the impact of mental health from other factors. The authors call for further research to understand the mechanisms at play and how workplace factors influence this relationship.

Finally, we looked into the study [4]. The study aims to investigate the economic burden of mental illness on workplace productivity. The data for this study comes from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. This is a large, ongoing survey that collects information from a representative sample of Australian households every year since 2001. The researchers used data from the first 13 waves (years 2001-2013) for their analysis of presenteeism and absenteeism.

The study found out that mental illness significantly reduces worker productivity, leading to economic losses. Overall, the study emphasizes the need for a multi-pronged approach to address the productivity losses associated with mental illness in the workplace. This includes reducing job stress, considering the complex interactions between job characteristics and mental health, and potentially developing targeted interventions for mentally ill workers.

3 Architecture

For our architecture we started by investigating the state of the art stack for **Smart Analysis in Big Data Systems** and selecting the technologies we felt were the most fitting. For the programming language we opted by **Python** since is the most adopted and package-wise complete for Data Science.

For the **database** we chose **CouchDB** instead of other technologies like HDFS or Cassandra since it has a high throughput, this will make our queries execute faster and retrieving the data in a more fluid form. Our datasets contain annual data so high write speeds into the database and handling huge files won't be the focus. As of the **data preparation/transformation** step we had two technologies in mind, **PySpark** and **Pandas**.

Regarding **data visualization** we chose **PowerBI** since it's the industry standard and has a connector to our database.

As of the architecture of the solution itself we had two approaches, ETL(Extract, Transform, Load) or ELT(Extract, Load, Transform)

In order to decide what would be the best/most fitting we developed both and after benchmarking them we will chose one.

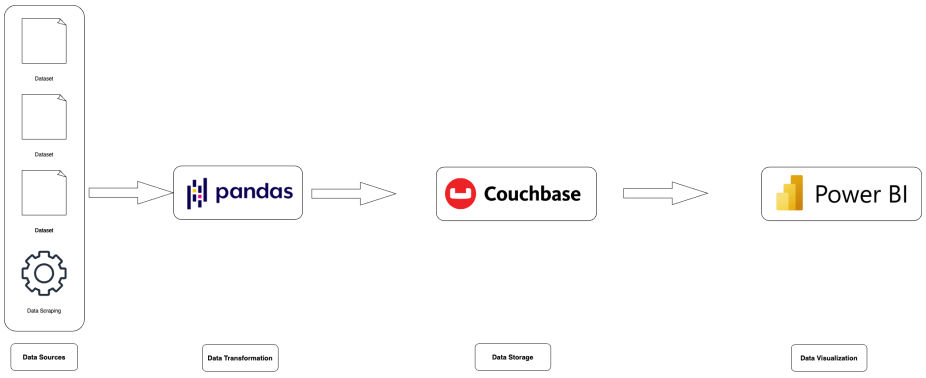


Figura 3.1: ETL architecture and the technologies used.

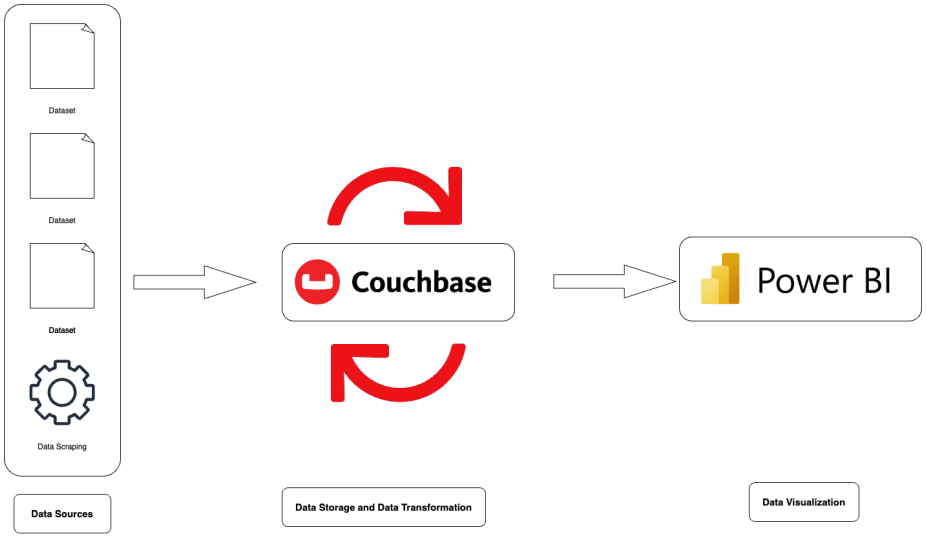
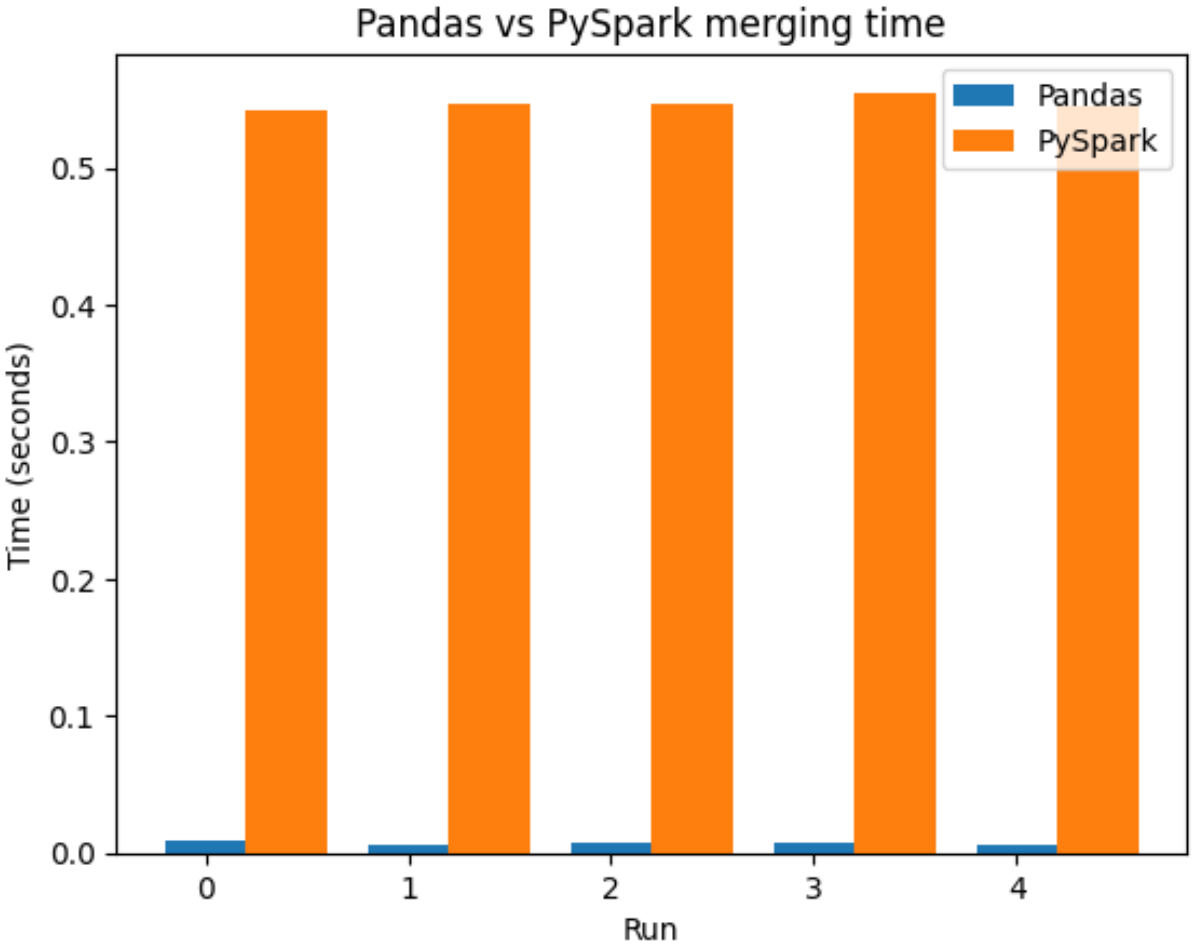


Figura 3.2: ELT architecture and the technologies used.

3.0.1 Data Transformation

For data transformation we have selected two of the most used technologies for Python that are Pandas and PySpark, in order to choose one we have benchmarked the time needed to merge the datasets with each one.



Despite PySpark being a dedicated tool for distributed data [15] processing, the benchmark results show that Pandas outperforms PySpark significantly in terms of merging datasets. The average time taken by Pandas for merging is approximately 0.0058 seconds, whereas PySpark takes around 0.5543 seconds on average. This considerable difference in performance suggests that for this specific task of merging these datasets, Pandas is a more efficient choice. However, it's important to note that PySpark may excel in other aspects of data processing, particularly when dealing with large datasets that cannot fit into memory which is not the case of our datasets. Therefore, the choice between Pandas and PySpark should be made based on the specific requirements and constraints of the data transformation tasks at hand.

3.0.2 Storage

Couchbase offers several advantages in terms of storage, making it a compelling choice for handling data in big data systems:

1. **High Performance:** Couchbase buckets store data persistently while also maintaining copies in memory, allowing for rapid access to frequently accessed data. This combination of disk and memory storage enhances performance, enabling faster retrieval of data compared to traditional databases.
2. **Flexible Ejection Policies:** Couchbase employs flexible ejection policies, particularly beneficial in distributed systems. These policies automate memory management when RAM quotas are exceeded,

eliminating the need for manual intervention from developers. This dynamic adaptation ensures optimal resource utilization across multiple machines.

3. **Bucket Management:** Couchbase allows for the creation of multiple buckets, each serving as a logical container for storing data. This approach enables effective segregation and management of data, providing better organization, security, and resource allocation. Additionally, buckets support various configurations and settings, allowing for fine-tuning based on specific use cases and requirements.

These features of Couchbase contribute to its effectiveness in big data systems by providing high performance, reliability, and flexibility in data storage and management. By leveraging these capabilities, we can achieve faster data retrieval, improved availability, and efficient resource utilization, enhancing overall system performance.

We chose Couchbase over alternatives like Cassandra or MongoDB due to its ability to leverage RAM for optimized data access and processing speed, aligning with our project goal of providing smooth and fast interaction with the graphical interface and data visualization step.

3.0.3 Storage Organization

Should we store a general dataset in the database with all the datasets merged or store them individually?

Our goal is to have fast queries in the database to ensure smooth and responsive visualization.

Analyzing the aforementioned approaches without any tests, storing the data in a general dataset may be ideal since we do not need to structure queries for each specific dataset. However, in some visualization cases, we may run queries on redundant data. On the other hand, storing data in separate datasets requires more effort, and structuring queries is also harder. However, in terms of maintainability, it may be better since the data from the data sources is filtered before being stored in each bucket of the database, and the queries may be more efficient since we can target them specifically to different buckets.

To confirm and verify our choice, we will benchmark both approaches.

3.0.4 Merging the datasets vs Keeping them separate

Before benchmarking the time to run the queries we can extract the following:

Keeping Datasets Separated	**Merging Datasets**
Pros:	**Pros:**
- Efficient for future insertions, especially if datasets are not updated simultaneously. - Granular control over data insertion and updates, potentially reducing contention and optimizing resource usage.	- Simplifies query writing and maintenance since there's only one dataset to query against. - Streamlines data access and may improve overall efficiency for certain query operations.
Cons:	**Cons:**
- May require more effort in query writing and maintenance due to multiple datasets. - Increased complexity in managing multiple datasets.	- Slightly slower execution times for certain queries compared to individual datasets. - Potential performance impact as the dataset grows, requiring careful monitoring and optimization.

Tabela 3.1: Pros and Cons of Keeping Datasets Separated vs. Merging Them

Keeping in mind the pros and cons of each approach we will write queries that resemble the activity that a user trying to visualize data would have, like the most improving country based on its productivity, the country with the lowest rates of mental illness and highest purchasing power, and average the times to run queries like the cited above over 5 runs.

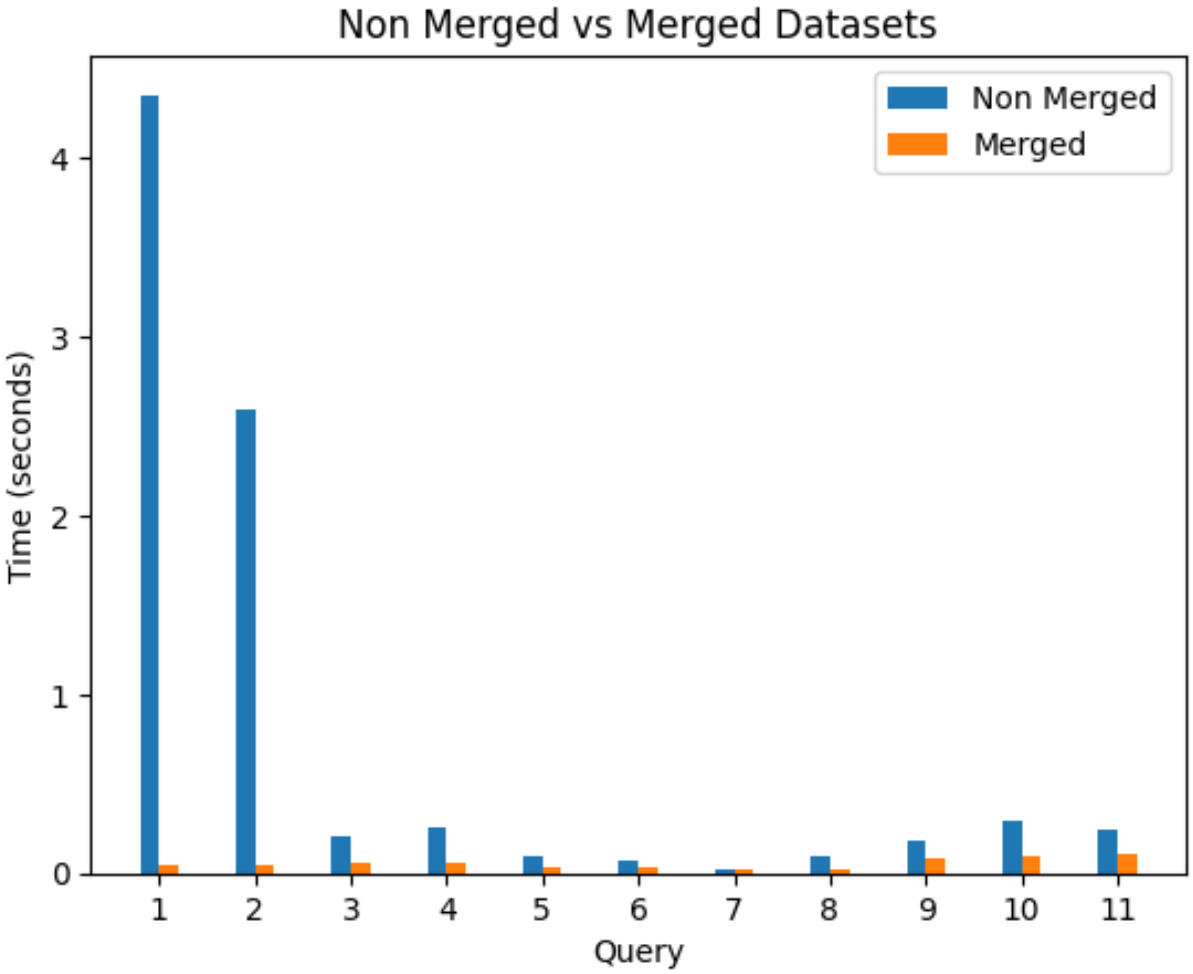


Figura 3.3: Average query execution times of the merged and non merged datasets over the course of 5 runs

The merged dataset queries outperform the separate dataset queries primarily due to the streamlined nature of data access. When datasets are merged, there's only one dataset to query against, simplifying the query writing and maintenance process. This streamlined approach reduces the overhead associated with managing multiple datasets and minimizes the need for complex joins or data retrieval operations across various datasets.

The merged dataset allows for optimized indexing and data organization, which can significantly enhance query performance. With all relevant data consolidated into a single dataset, the database can efficiently execute queries without the overhead of joining disparate datasets on the fly despite the indexes created in the separate datasets.

This results will facilitate all the process of writing the queries but significantly dificultate the process of asynchronous data insertion in the database.

3.0.5 Architecture Type

As mentioned before, there are two possible architectures for our solution, ETL (Extract, Transform, Load) or ELT (Extract, Load, Transform).

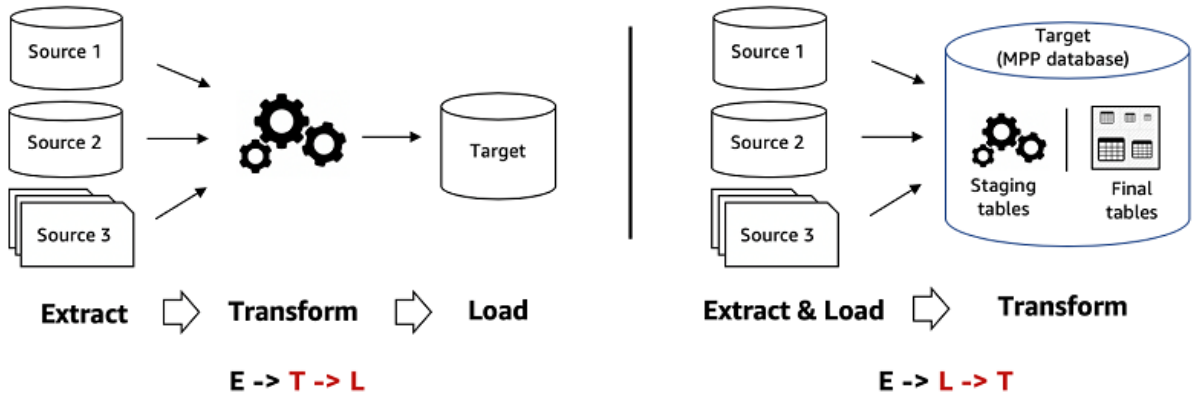


Figura 3.4: ETL vs ELT

What are the key differences between the ETL and ELT?[2]

ETL process

ETL has three steps:

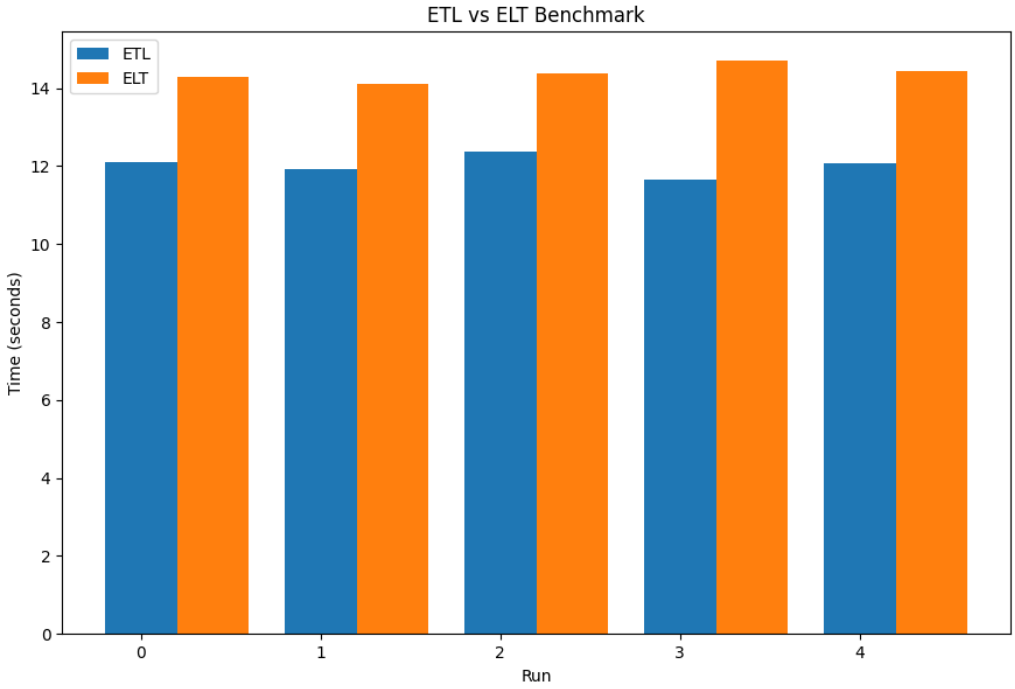
- You extract raw data from various sources
- You use a secondary processing server to transform that data
- You load that data into a target database
- The transformation stage ensures compliance with the target database's structural requirements.
- You only move the data once it is transformed and ready.

ELT process

These are the three steps of ELT:

- You extract raw data from various sources
- You load it in its natural state into a data warehouse or data lake
- You transform it as needed while in the target system
- With ELT, all data cleansing, transformation, and enrichment occur within the data warehouse.
- You can interact with and transform the raw data as many times as needed.

The extraction part is the same for both approaches, but the data transformation and it's storage may differ,



3.0.6 Final Architecture for our Solution

Our final architecture for the solution is based on the Extract, Transform, Load (ETL) approach with a merged dataset. This architecture is depicted in the figure below:

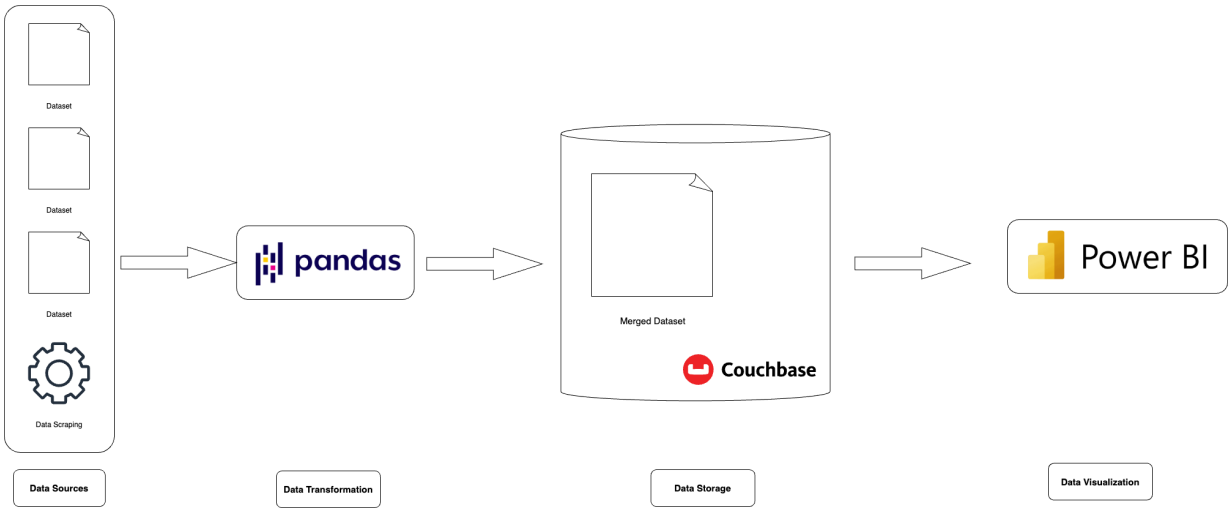


Figura 3.5: Final Architecture

Key components of our architecture include:

- **Data Extraction:** Raw data is extracted from various sources like yearly published datasets and scraped.

- **Data Transformation:** The extracted data is then transformed using **Pandas** and merged to optimize performance.

- **Data Loading:** The transformed dataset is loaded into its target **Couchbase** bucket for storage.

- **Data Visualization:** The stored data will be available through **PowerBI** using the provided connector to our database.

3.1 Data analysis

In order to fulfill the goal of our project, we selected and used 4 datasets. We made sure to find data from World Renowned Sources to maintain data integrity and validity. All of the found data came in a CSV format. Further down the road, in the data preparation/cleaning process, we used an API to convert the salary data to USD instead of the country's currency. We should note that not all datasets have the same time span, therefore some of the data (mostly historical) has been omitted.

3.1.1 Cost of Living

This dataset, retrieved from [10], via a scraping script developed by us, provides information related to the cost of living in 158 different countries from 2009 up until 2023. The cost of living provided on this dataset are relative to New York City, with a baseline index of 100% for New York City. [9] It contains these features:

- **Date** : the year where the data came from.
- **Country** : the country where the data came from.
- **Cost of Living** : the index indicates the relative prices of consumer goods like groceries, restaurants, transportation, and utilities. It excludes accommodation expenses such as rent or mortgage. For instance, a city with a Cost of Living Index of 120 is estimated to be 20% more expensive than New York City (excluding rent). [9]
- **Rent Index**: the index estimates the prices of renting apartments in a city compared to New York City. [9]
- **Cost of Living Plus Rent Index** : the index estimates consumer goods prices, including rent, in comparison to New York City.[9] .
- **Groceries Index** : the index provides an estimation of grocery prices in a city relative to New York City. [9]
- **Restaurant Price Index** : this index compares the prices of meals and drinks in restaurants and bars to those in New York City. [9]
- **Local Purchasing Power Index** : this index indicates the relative purchasing power in a given city based on the average net salary. A domestic purchasing power of 40 means that residents with an average salary can afford, on averagem 60% less goods and services compared to residents of New York City with an average salary. [9]

3.1.2 Mental Illness

This dataset, retrieved from [14], provides information related to the number of individuals that suffer from mental illness diseases in 228 countries from 1990 up until 2019. The features in the dataset are:

- **Entity** : country where the data came from.
- **Code** : code that represents the country where the data came from.
- **Year** : year where the data came from.

In order to explain the following features we must know the definition of DALYs which is Disability-Adjusted Life Years are a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability, or early death. Given that we have:

- **DALYs from depressive disorders per 100,000 people in, both sexes aged age-standardized**: refers to the average number of years of healthy life lost due to depressive disorders per 100,000 individuals in a population, adjusted for age differences and accounting for both males and females.
- **DALYs from schizophrenia per 100,000 people in, both sexes aged age-standardized**: refers to the average number of years of healthy life lost due to schizophrenia per 100,000 individuals in a population, adjusted for age differences and accounting for both males and females.
- **DALYs from bipolar disorder per 100,000 people in, both sexes aged age-standardized**: refers to the average number of years of healthy life lost due to bipolar disorder per 100,000 individuals in a population, adjusted for age differences and accounting for both males and females.
- **DALYs from eating disorders per 100,000 people in, both sexes aged age-standardized**: refers to the average number of years of healthy life lost due to eating disorders per 100,000 individuals in a population, adjusted for age differences and accounting for both males and females.
- **DALYs from anxiety disorders per 100,000 people in, both sexes aged age-standardized**: refers to the average number of years of healthy life lost due to anxiety disorders per 100,000 individuals in a population, adjusted for age differences and accounting for both males and females.

3.1.3 Wold Annual Wage

This dataset, retrieved from [12], provides information about the annual age in 39 different countries. The features presented in the dataset are:

- **Country**: country where the data came from.
- **Year**: year where the data came from.
- **Unit Code** : code that represents the money unit of the annual wage. There's 22 money units presented in the dataset.
- **Unit** : the money unit of the annual wage.

3.1.4 World Labor Productivity

This dataset, retrieved from [8], provides information about the productivity of 69 countries from 1950 up until 2019. In this context, the productivity is measured given the GDP (output, multiple price benchmarks) divided by the Annual working hours per worker and the Number of people in work.

- **Entity** : country where the data came from.
- **Code** : code that represents the country where the data came from.
- **Year** : year where the data came from.

- **Productivity (output per hour worked)** : GDP divided by the Annual working hours per worker and the Number of people in work

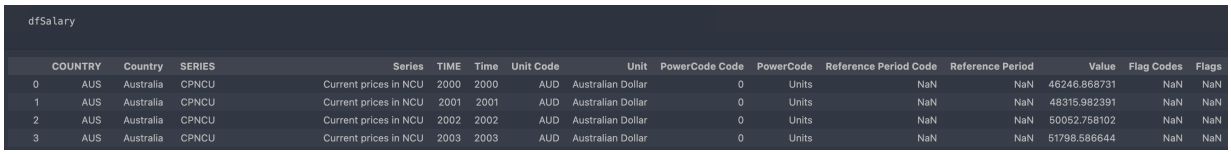
As justified before our architecture will consist in an ETL approach, using Couchbase for the data warehouse and pandas for the transformation and merge of the datasets.

3.2 Extraction

For the data exctraction process we downloaded all the CSV files from the sources, except the Salary data. For the salary data we wrote a *Python* script that *scrapped* the data from the website and stored it in a CSV in order to not overload the source website with requests in the course of the work. With all the data collected we delved into the next step of the pipeline.

3.3 Transformation

For the transformation of the dataset we started by analysing each dataset in order to verify the missing data, negligible and out of context data.



dfSalary

	COUNTRY	Country	SERIES	Series	TIME	Time	Unit Code	Unit	PowerCode Code	PowerCode	Reference Period Code	Reference Period	Value	Flag Codes	Flags
0	AUS	Australia	CPNCU	Current prices in NCU	2000	2000	AUD	Australian Dollar	0	Units	NaN	NaN	46246.868731	NaN	NaN
1	AUS	Australia	CPNCU	Current prices in NCU	2001	2001	AUD	Australian Dollar	0	Units	NaN	NaN	48315.982391	NaN	NaN
2	AUS	Australia	CPNCU	Current prices in NCU	2002	2002	AUD	Australian Dollar	0	Units	NaN	NaN	50052.758102	NaN	NaN
3	AUS	Australia	CPNCU	Current prices in NCU	2003	2003	AUD	Australian Dollar	0	Units	NaN	NaN	51798.586644	NaN	NaN

Figura 3.6: Example of data to be cleaned

3.3.1 Data removal

As inspected above, there are several columns with *NaN* values only and we wont need them for our analysis so we will drop them.

We also dropped some columns that we believed that weren't significant for our research such as:

- **'Code'** feature from the Labour dataset.
- Some values that were duplicate or non relevant in the Salary dataset: **'COUNTRY' , "Flag Codes", "Flags", "SERIES", "Series", "Unit", "PowerCode", "PowerCode Code","Reference Period Code", "Reference Period"**

3.3.2 Renaming of attributes for normalization

In order to normalize querying and analysis through all the data sources, we had to rename for instance "Entity"to "Country"in Mental Health and Labor datasets.

Some datasets also had the feature "Year"as "Time", so we renamed them to "Year".

3.3.3 Currency Exchange

It was determined that the data from the Wage dataset came in each countries currency. In order to normalize the data, we elaborated a function that based on the dataset would get each country exchange rate by year and we created a dataset to store this data. We then applied these exchange rates in order to have all the salaries in USD.

After all the previous cleanup we followed up by proceeding to merge all the datasets. As analysed previously, queries made to the merged dataset are significantly faster than the ones made to individual tables in the database.

3.4 Loading

Upon finishing the merge of the datasets we proceeded to load the each countries information to our *Couchbase* instance.

Each countries document is as follows:


```
{
  "Year": "30/12/2019",
  "Country": "Switzerland",
  "CoL": 121.2,
  "Rent Index": 50.2,
  "Cost of Living Plus Rent Index": 87.1,
  "Groceries Index": 120.8,
  "Restaurant Price Index": 123.1,
  "Local Purchasing Power Index": 129.7,
  "Productivity": 82.918655,
  "Unit Code": "CHF",
  "Salary": 80501.2142838992,
  "Exchange Rate": 1.101884,
  "Code": "CHE",
  "DALYs from depressive disorders per 100,000 people": 635.3109,
  "DALYs from schizophrenia per 100,000 people": 178.6776,
  "DALYs from bipolar disorder per 100,000 people": 202.4897,
  "DALYs from eating disorders per 100,000 people": 107.11581,
  "DALYs from anxiety disorders per 100,000 people": 650.5403
}
```


3.5 Data Visualization

In order to visualize the data we have got so far we resorted to PowerBI as the software.

To establish the communication between the visualization software and our database we used the PowerBI Couchbase connector.

At this stage we already have the transformed data in PowerBI therefore the only task left is to create dashboards that can translate to valuable insights regarding our problem. The first dashboard that we created aimed to represent the *global* perspective of certain parameters such as each countries salary, the impact of the salary in the country productivity, the country with the highest number of Mental Illnesses per 100000 inhabitants and the country with the highest cost of living.

Aiming to create more insightful visualizations we elaborated a metric that dictates the quality of life in each country by year taking into consideration some of the following metrics published by the World Health Organization(WHO): *Standard indicators of the quality of life include wealth, employment, the environment, physical and mental health, education, recreation and leisure time, social belonging, religious beliefs, safety, security and freedom.*

Our metric that takes into account the cost of living index, average salary(Wealth), prevalence of mental

illnesses per 100,000 inhabitants(Mental Health), and the restaurant price index(Leisure).

This metric will help us *rank* the countries by the impact of the cost of living, salary and mental health.

The dashboards we have created are the following:

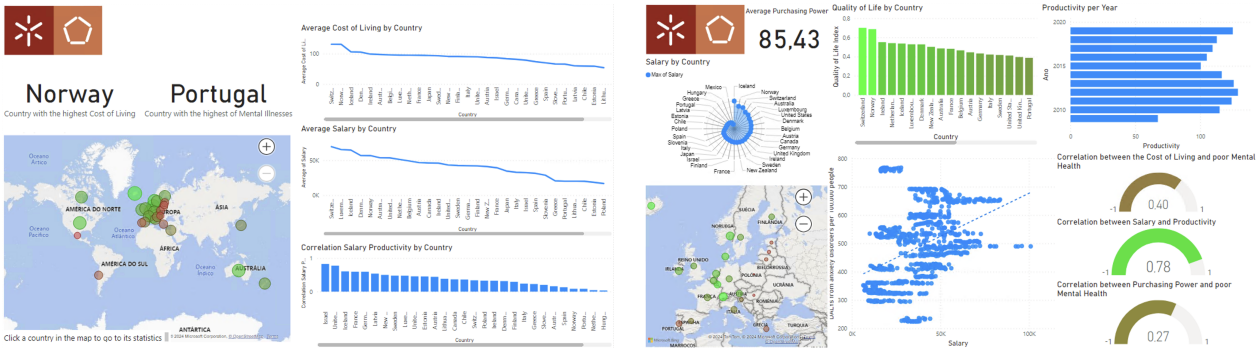


Figura 3.7: Global Dashboards

The figure above shows some general metrics and allows us to evaluate that Norway is the country with the highest cost of living, whilst Portugal is the country with the highest occurrences of mental illnesses per citizen.

In order to analyse some case studies we ended up choosing the data for the United States of America and Norway, since Norway is the country with the highest cost of living, and the USA since we have verified that is one of the countries with the highest cost of living and occurrences of mental illnesses.

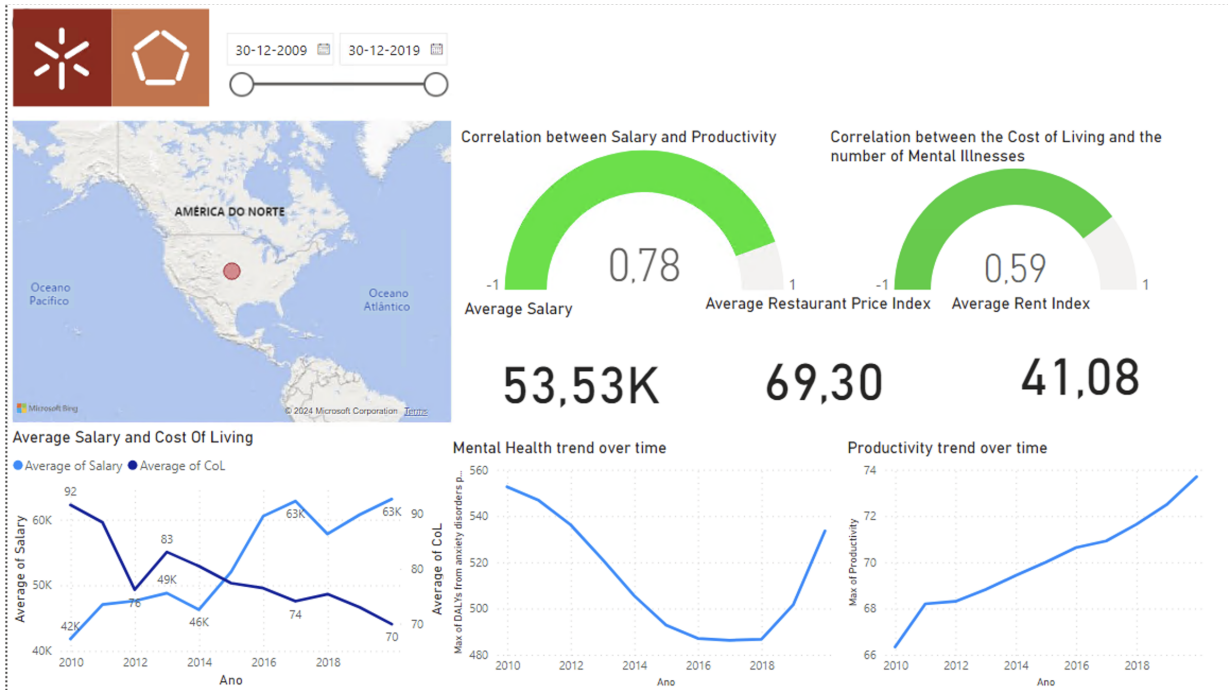


Figura 3.8: USA case study from 2009 to 2019

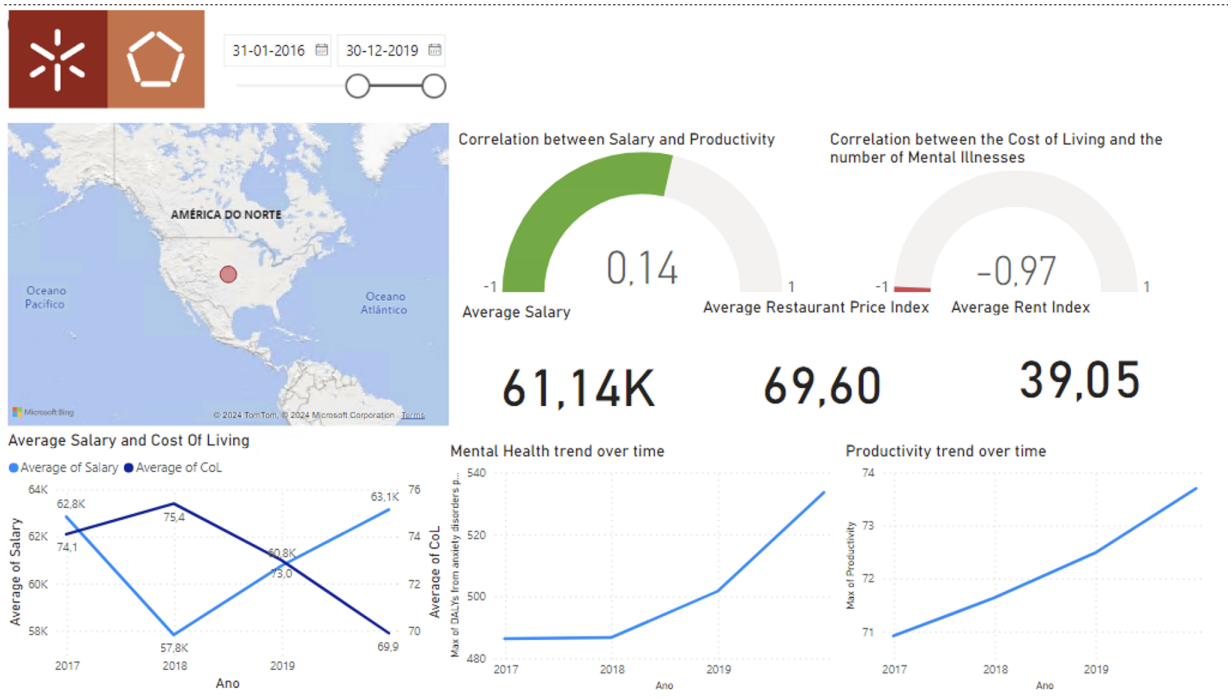


Figura 3.9: USA case study from 2016 to 2019

It is clear that, in a global view (considering all the data), the cost of living is strongly positively correlated with the number of mental illnesses in the USA. However, if we restrict the analysis to the years 2016 to 2019, we observe a strong negative correlation between the cost of living and the number of mental illnesses. This means that despite the lowering of the cost of living during this period, the number of mental illnesses continued to rise. This trend could be attributed to other factors, such as political influences, as cited by the ***NIH***[3].

We can also extract from the dashboards above that throughout the years, the salary and productivity have always been strongly correlated. More insights and possible conclusions can be extracted with the analysis of the cases of other countries.

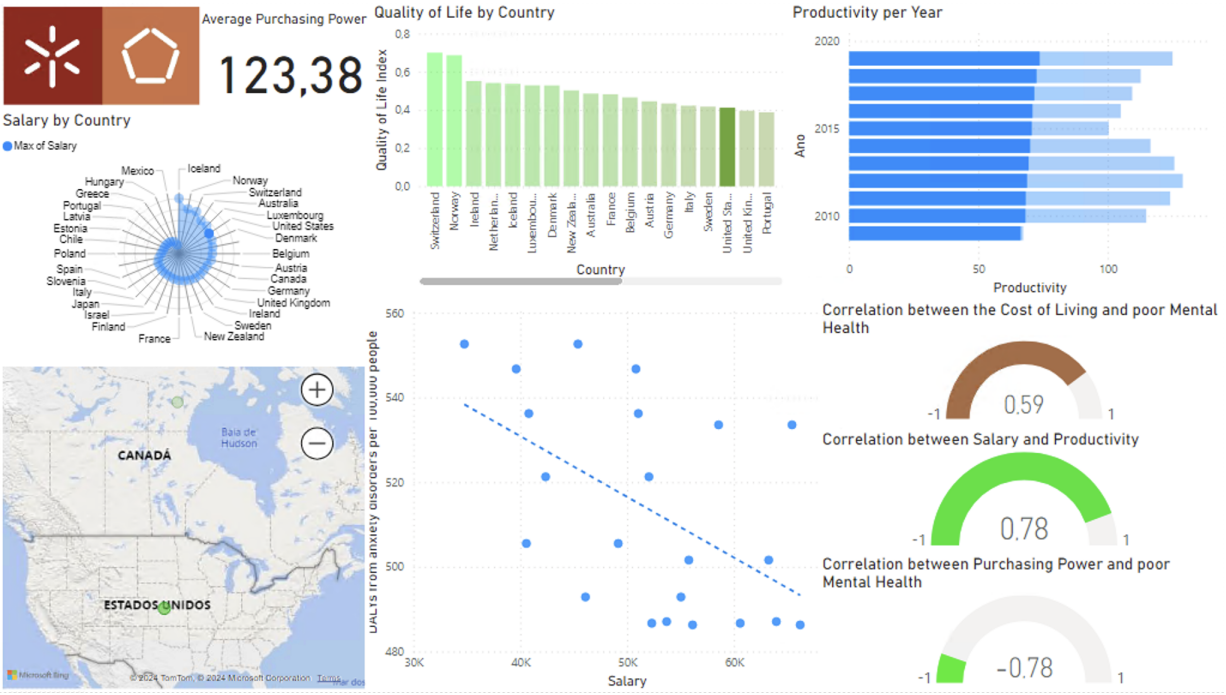


Figura 3.10: USA global ranking dashboard

The dashboard above clearly illustrates that regarding our metric of *Quality of Life*, ranks 16th in our dataset and that besides the year 2009 the United States aren't the amongst most productive countries. We can also see that the purchasing power is strongly negatively correlated with the number of mental illnesses therefore, the purchasing power may play a strong role in the well being of the USA inhabitants.

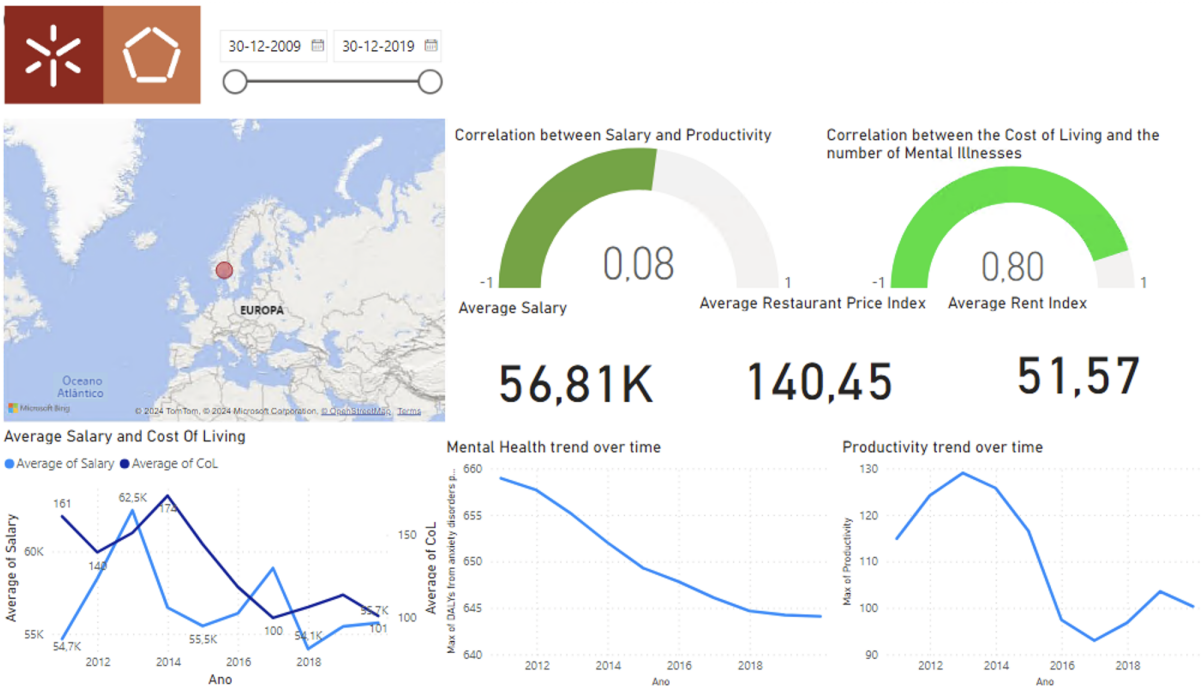


Figura 3.11: Norway case study from 2010 to 2019

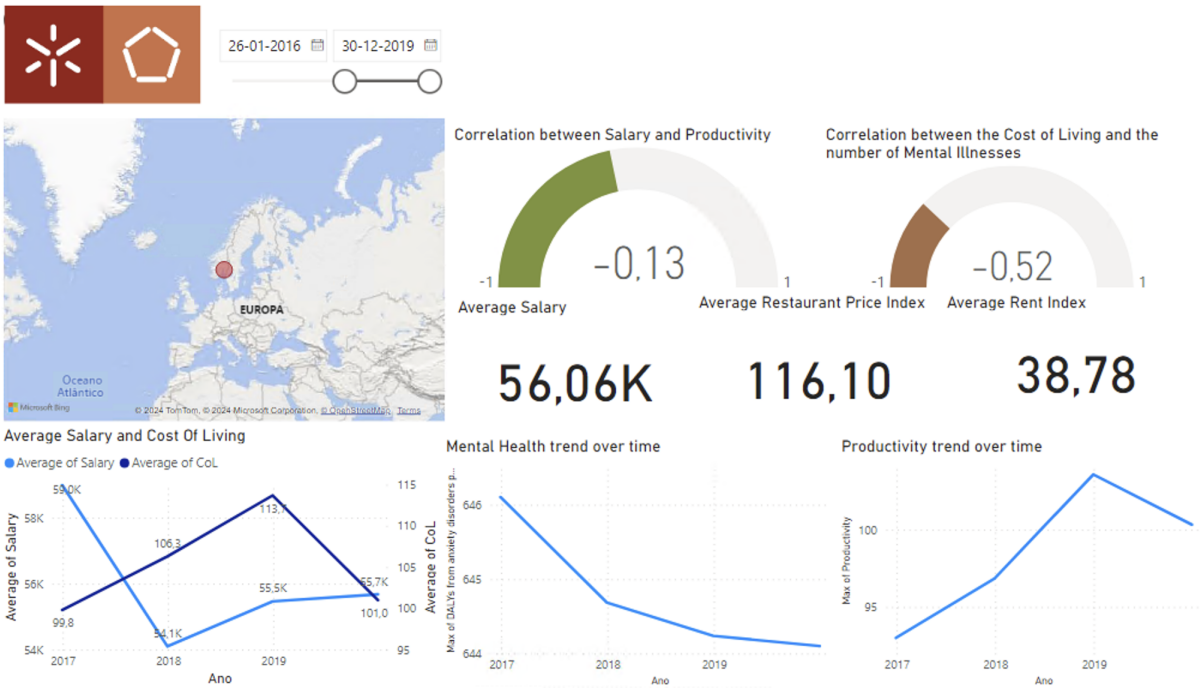


Figura 3.12: Norway case study from 2016 to 2019

The figure 3.12 illustrates a positive correlation between salary and productivity in Norway between 2010 and 2019. While the correlation is not extremely strong, with a coefficient of 0.08, it suggests that as salary increases, productivity tends to rise as well.

It also shows that in contrast, the correlation between the cost of living and the number of mental illnesses is significantly stronger. With a coefficient of 0.80, the relationship is almost linear, indicating a strong and direct association between the two variables. This metric is likely to be the most important factor in understanding the relationship between these variables.

The figure 3.13 shows the scenario of Norway between 2016 and 2019. The analysis of Norwegian data between 2016 and 2019 reveals a complex interplay between salary, productivity, and cost of living. While the correlation between salary and productivity is surprisingly negative, with a coefficient of -0.13, this suggests that as salary increases, productivity tends to decrease. This finding might seem uncanny but it could be due to increased salaries being offset by increased costs of living.

Conversely, the correlation between the cost of living and mental illnesses occurrences exhibits a strong negative relationship, with a coefficient of -0.52. This implies that as the cost of living rises, the number of mental illnesses occurrences tends to decrease. Although it might seem weird, in our opinion this could be due to increased access to resources and services that help alleviate mental health issues or the existence of a "tipping point"where high costs of living lead to a collective sense of resilience and adaptability. In the articles [7] and [6] we can find some information that might justify the weird findings above.

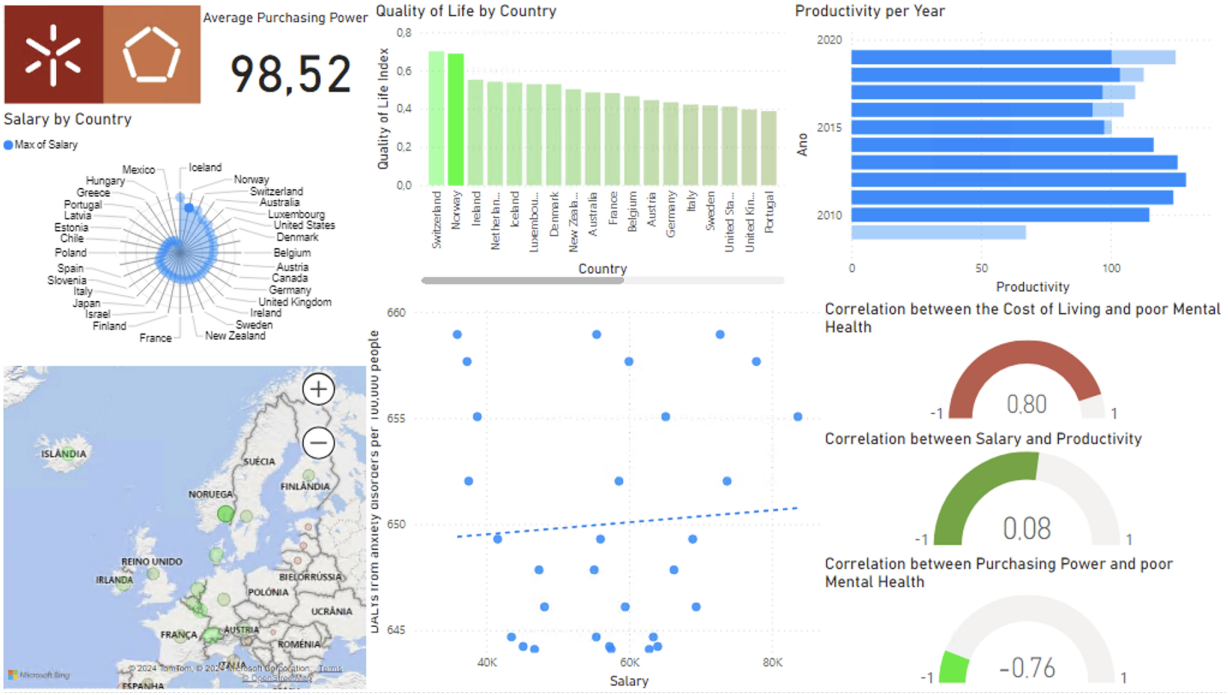


Figura 3.13: Norway global ranking dashboard

In the dashboard above we can assert that despite the high cost of living Norway ranks second in our *Quality of Life* metric. We can also notice a clear decrease in the productivity from 2014 on, this can be due to ***sustainability*** problems[11]. We can confirm that just like the USA case study, the Purchasing Power of the population has a strong correlation with the Mental Health of the countries inhabitants.

As a benchmark we will also analyse the Portuguese data since it is the country with the highest index of mental illnesses.

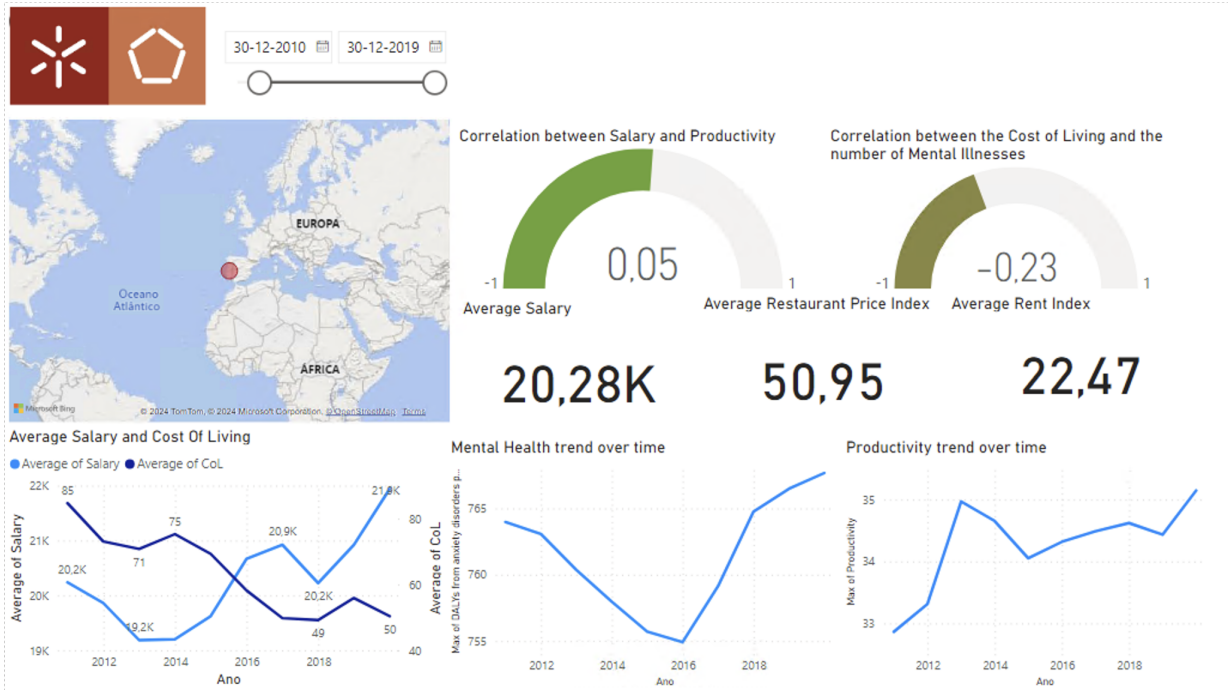


Figura 3.14: Portugal case study from 2010 to 2019

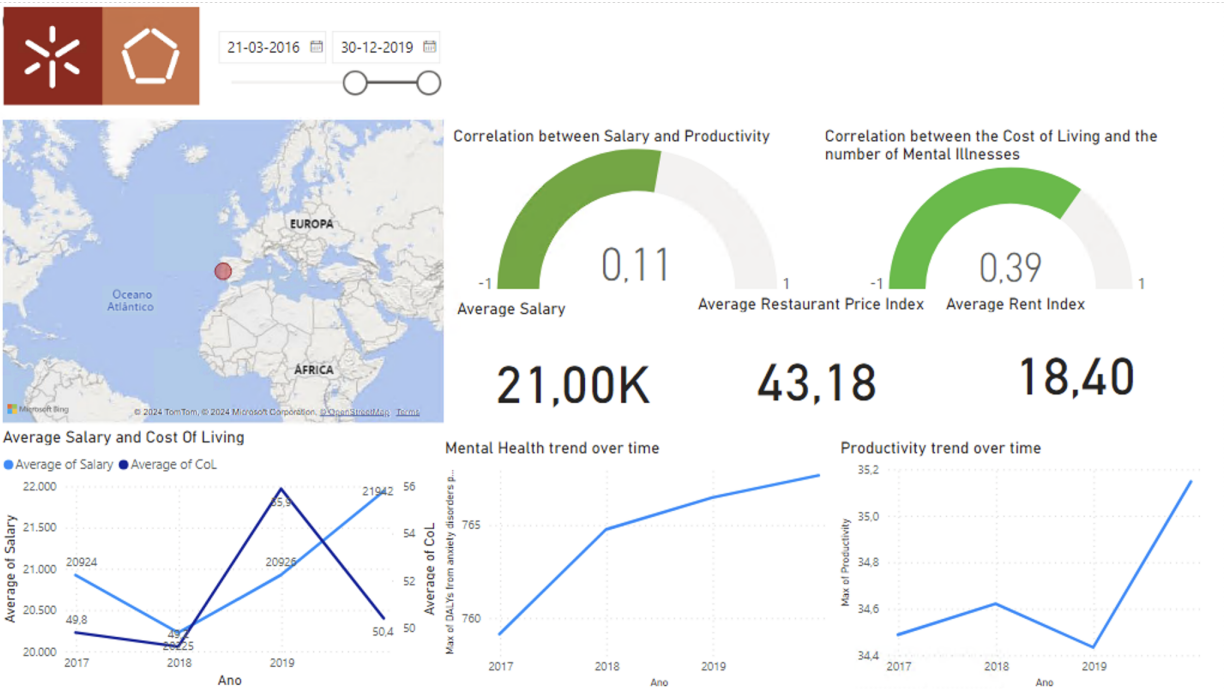


Figura 3.15: Portugal case study from 2016 to 2019

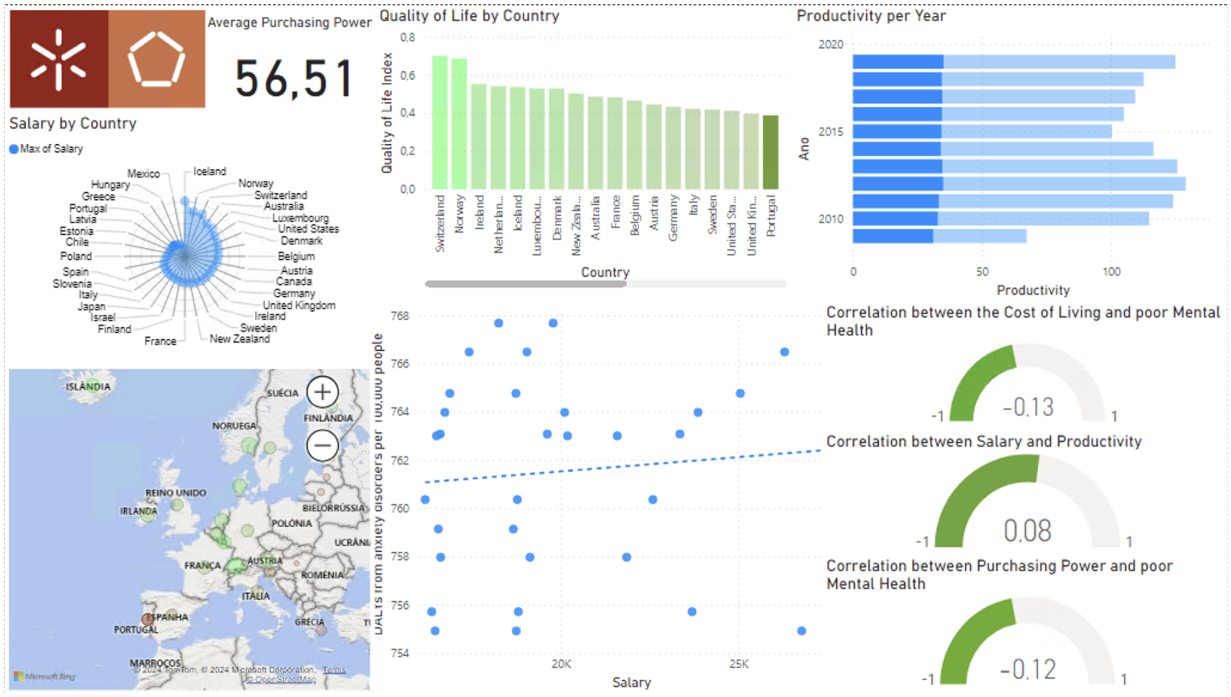


Figura 3.16: Portugal global ranking dashboard

In the dashboard above, we can infer that Portugal has a below-average purchasing power, it has the highest index of inhabitants with poor mental health, and its productivity has been consistently below average. We can also infer that in our dataset, Portugal's ranking is only two places below the United States, a country with high productivity and high purchasing power, this can indicate that are another factors that play a strong role in the Portuguese quality of life, upon further analysis, we noticed that Portugal has the eighth lowest cost of living, this factor can *counter-balance* the low salaries and high number of mental illnesses amongst the Portuguese population.

We can also verify that Portugal has a negligible correlation between salary and productivity. Globally, the wage has a strong direct correlation to productivity. This suggests that other factors, such as ***the qualifications of the Portuguese Managers***[13], economic policies or cultural aspects, might play a significant role in the Portuguese productivity.

4 Conclusion

In this report, we investigated the complex relationship between purchasing power, mental health, and productivity across various countries using a robust big data architecture and advanced analytical techniques. Our comprehensive approach encompassed the selection, preparation, integration, and analysis of multiple datasets, followed by the visualization of the resulting insights.

The architecture for our solution was carefully chosen to optimize performance and efficiency. We adopted an ETL (Extract, Transform, Load) approach, utilizing Couchbase for data storage and Pandas for data transformation. This setup allowed for rapid data retrieval and streamlined data manipulation, essential for generating timely and accurate insights.

Our analysis revealed several significant findings:

- The cost of living has a varying impact on mental health and productivity depending on the country and time period. For instance, in the USA, we observed a strong positive correlation between the cost of living and mental health issues in a global view, while a more granular analysis (2016-2019) indicated a negative correlation.
- Norway, despite having the highest cost of living, demonstrated a strong positive correlation between salary and productivity from 2010 to 2019. However, from 2016 to 2019, the data suggested a complex relationship where rising costs of living correlated with a decrease in mental health issues.
- Portugal presented a unique case with its low cost of living seemingly mitigating the negative impacts of low purchasing power and high mental health issues. The negligible correlation between salary and productivity highlighted the influence of other factors, such as managerial qualifications and economic policies, on productivity.

The visualizations created using PowerBI provided a clear and interactive way to understand these complex relationships. The dashboards illustrated key metrics and trends, enabling a deeper understanding of how purchasing power impacts mental health and productivity in different countries.

In conclusion, our study underscores the importance of considering multiple socioeconomic factors when analyzing the well-being and economic performance of countries. The strong correlations identified in our analysis suggest that interventions aimed at improving purchasing power could have significant positive effects on mental health and productivity. However, the nuanced differences between countries indicate that tailored approaches are necessary. Future research could further explore these relationships, taking into account additional variables and broader datasets to build on our findings.

This report contributes to the growing body of knowledge on the intersection of economics, health, and productivity, demonstrating the power of big data analytics in uncovering critical insights and informing policy decisions.

5 Bibliografia

[1] Afis A Agboola, Oluwaseun T Esan, Oluwasegun T Afolabi, Taiwo A Soyinka, Adedunmola O Oluwaranti, and Adeniji Adetayo. Economic burden of the therapeutic management of mental illnesses and its effect on household purchasing power. *PloS one*, 13(9):e0202396, 2018.

[2] AWS. What's the difference between etl and elt? `https://aws.amazon.com/compare/the-difference-between-etl-and-elt/`. Accessed: March 25, 2024.

[3] corresponding author1 Renee Y. Hsia MD MSc 1 2 Victoria Yeung 3 Brandon W. Yan, BA and PhD4 Frank A. Sloan. Changes in mental health following the 2016 presidential election. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7602772/`, 2021.

[4] Melisa Bubonya, Deborah A Cobb-Clark, and Mark Wooden. Mental health and productivity at work: Does what you do matter? *Labour economics*, 46:150–165, 2017.

[5] Claire de Oliveira, Makeila Saka, Lauren Bone, and Rowena Jacobs. The role of mental health on workplace productivity: a critical review of the literature. *Applied health economics and health policy*, 21(2):167–193, 2023.

[6] Organisation for Economic Co-operation and Development. Productivity. 2023.

[7] Arne Holte. Promotion of mental health and prevention of mental disorders in a rich welfare state: A norwegian perspective. *Mental Health & Prevention*, 33:200321, 2024.

[8] Our World in Data. Labor productivity per hour, penn world table. `https://ourworldindata.org/grapher/labor-productivity-per-hour-pennworldtable?tab=table`, Accessed on 14 March 2024.

[9] Numbeo. Cost of living - consumer price index (cpi) explained. `https://www.numbeo.com/cost-of-living/cpi_explained.jsp`. Accessed: March 14, 2024.

[10] Numbeo. Cost of living rankings by country. `https://www.numbeo.com/cost-of-living/rankings_by_country.jsp`, Accessed on 14 March 2024.

[11] OECD. Norway : Wellbeing is high, but must be sustained. `https://www.oecd-ilibrary.org/sites/c217a266-en/index.html?itemId=/content/publication/c217a266-en`, 2019.

[12] Organisation for Economic Co-operation and Development. Average annual wages. `https://stats.oecd.org/index.aspx?DataSetCode=AV_AN_WAGE`, Accessed on 14 March 2024.

[13] European Union. The productivity of the portuguese economy. `https://economy-finance.ec.europa.eu/system/files/2019-11/pt2019`, 2019.

[14] World Health Organization. Mental health. `https://www.who.int/health-topics/mental-health#tab=tab_1`, Accessed on 14 March 2024.

[15] Ahmed Uz Zaman. Pandas vs pyspark. https://medium.com/geekculture/pandas-vs-pyspark-fe110c266e5c. Accessed: March 25, 2024.