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

Ana Rita Poças pg53645 and Nelson Almeida pg52697

Department of Informatics, University of Minho , Rua da Universidade, 4710-057, Braga, Portugal.

Contributing authors: pg53645@alunos.uminho.pt; pg52697@alunos.uminho.pt;

Abstract

This paper presents the process and analysis undertaken for a practical assignment in the "Análise Inteligente em Sistemas de Big Data" (Intelligent Analysis in Big Data Systems) course. The assignment involved selecting correlated datasets, performing meticulous data preparation, merging the datasets into a comprehensive database, and finally visualizing the integrated data to extract valuable insights.

Keywords: big data, cost of living, mental health, productivity, world annual wage

1 Introduction

This paper details the process and analysis involved in completing a practical assignment for the Análise Inteligente em Sistemas de Big Data (Intelligent Analysis in Big Data Systems) course. The assignment centered on selecting correlated datasets, meticulously preparing the data for integration, merging them into a comprehensive database, and finally visualizing the combined data to extract valuable insights.

We have chosen to analyse the impact of the purchasing power in mental health and productivity of different countries and for that we began by choosing datasets that could provide the data that we needed for the analysis.

2 State of 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 [2] 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 [3]. 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 Datasets

The datasets chosen for our study are the following:

3.1 Cost of Living

This dataset, retrieved from [4], 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. [5] 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). [5]
- **Rent Index**: the index estimates the prices of renting apartments in a city compared to New York City. [5]
- **Cost of Living Plus Rent Index** : the index estimates consumer goods prices, including rent, in comparison to New York City.[5] .
- **Groceries Index** : the index provides an estimation of grocery prices in a city relative to New York City. [5]
- **Restaurant Price Index** : this index compares the prices of meals and drinks in restaurants and bars to those in New York City. [5]
- **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. [5]

3.2 Mental Illness

This dataset, retrieved from [6], 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.

3.3 Wold Annual Wage

This dataset, retrieved from [7], 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.

3.4 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

4 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.

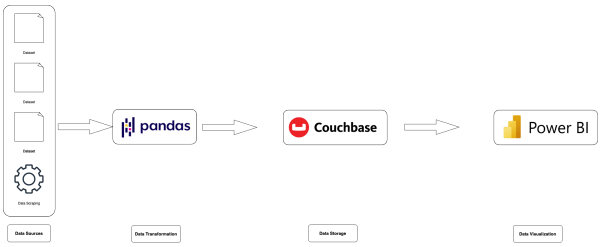

Fig. 1: ETL architecture and the technologies used.

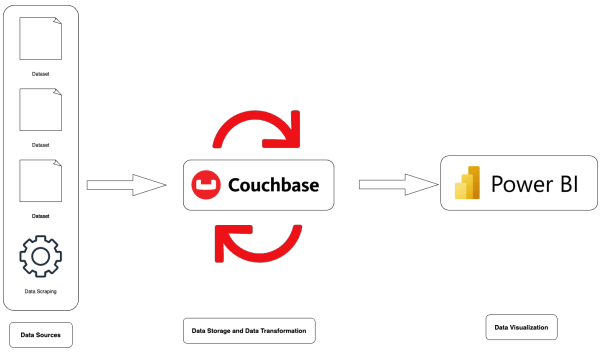

Fig. 2: ELT architecture and the technologies used.

4.1 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.

Run	Pandas	PySpark
1	0.0079	0.5563
2	0.00533	0.5495
3	0.0047	0.5730
4	0.0063	0.5496
5	0.0051	0.5433
Average	0.0058	0.5543

Table 1: Pandas and PySpark Performance Metrics

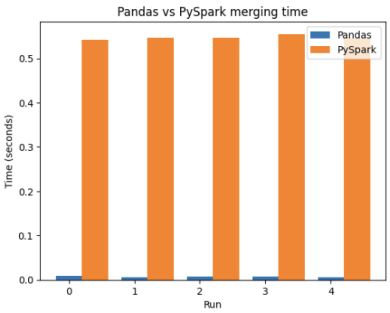

Fig. 3: Pandas vs PySpark performance benchmark

Despite PySpark being a dedicated tool for distributed data [9] processing, the benchmark results show that Pandas outperforms PySpark significantly in terms of merging our 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 larger 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.

4.2 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 stores data on disk for persisntence but also keeps copies in memory for frequently accessed data. This combination allows for much quicker retrieval compared to traditional databases.
2. **Flexible Ejection Policies:** Couchbase automatically manages memory in distributed systems by ejecting less-used data when RAM fills up. This eliminates manual intervention and optimizes resource usage across machines.
3. **Bucket Management:** Couchbase uses buckets to logically group and manage data, improving organization, security, and resource allocation. These buckets can be customized for specific needs.

These features of Couchbase contribute to its effectiveness in big data systems thanks to its high-perfomance data management. In memory storage and flexible memory management features contribute to faster data acess, reliable operation, and efficient resource use.

We chose Couchbase over alternatives like Cassandra or MongoDB due to the fact that Couchbase's focus on RAM optimization for data access aligns well with the need for speed in data visualization.

4.3 Storage Organization

To ensure fast queries and smooth data visualization, we evaluated two approaches for storing data in the database: merging all datasets or keeping them separate.

Initially, merging datasets seemed ideal for faster querying without needing specific structures for each one. However, this could lead to redundant data impacting query efficiency in some visualizations. Separate datasets, although requiring more upfront effort and complex queries, offer potential benefits in maintainability – data is pre-filtered, and queries can target specific buckets for better efficiency.

To determine the most efficient method, we'll benchmark both merged and separate data storage for visualization.

4.4 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.

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.

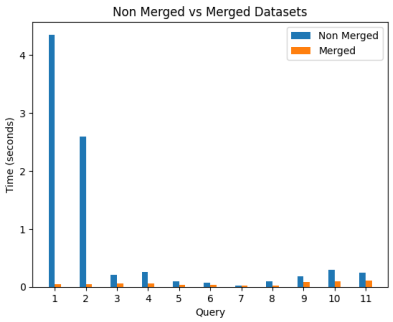

Fig. 3: Average query execution times of the merged and non merged datasets over the course of 5 runs

Merging datasets demonstrably enhances query performance and manageability. This approach simplifies the query development process by eliminating the need to join and integrate data from multiple sources. A unified data repository facilitates streamlined data access, minimizing the overhead associated with managing and querying disparate datasets. Consequently, merged datasets enable the formulation of more concise and efficient queries and it enables optimized indexing, streamlining query execution and boosting performance.
While merged datasets simplify query writing and potentially improve performance, they can complicate asynchronous data insertion (adding data at different times).

4.5 Architecture Type

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

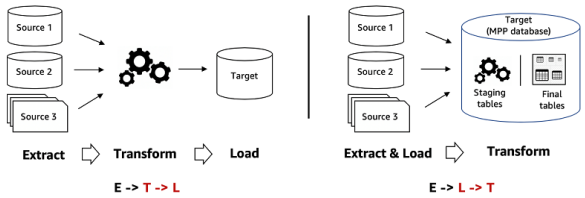

Fig. 4: ETL vs ELT

ETL Approach	ELT Approach
1. Extracting raw data from various sources 2. Transforming data using a secondary processing server 3. Loading transformed data into a target database	1. Extracting raw data from various sources 2. Loading raw data into a data warehouse or data lake 3. Transforming data within the target system

Table 1: Comparison of ETL and ELT Approaches

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

Run	ETL	ELT
1	11.7082	14.5982
2	11.7456	14.8040
3	12.3647	14.6681
4	12.2182	14.6410
5	12.0346	14.8606
Average	12.014	14.7143

Table 3: ETL vs ELT Performance Metrics

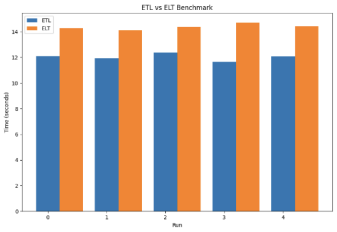

Fig. 6: ETL vs ELT performance benchmark

4.6 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:

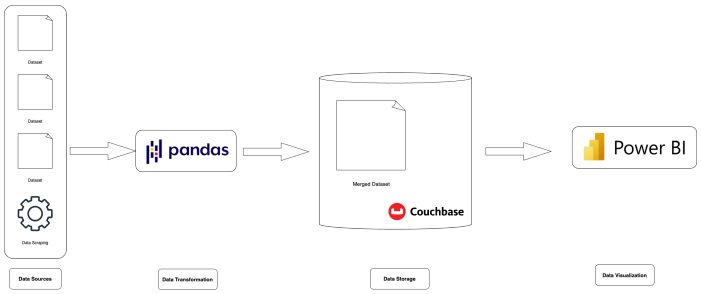

Fig. 5: Final Architecture

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.

5 Results

In this section, we will present the dashboards created in Power BI to visualize the data in order to answer our case study.

Aiming to summarize our work 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).

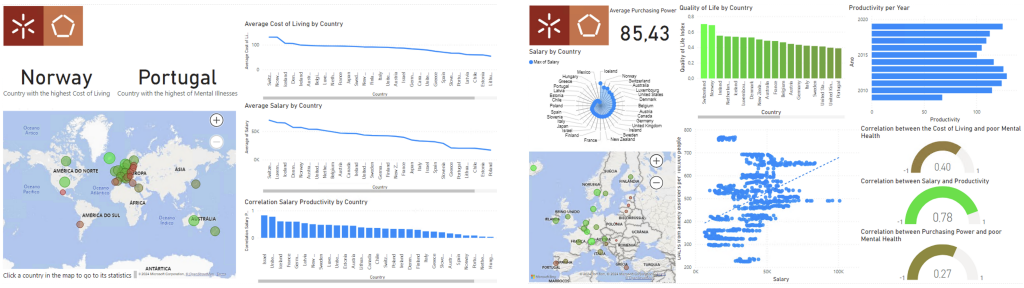

Fig. 6: Global dashboards

Case studies:

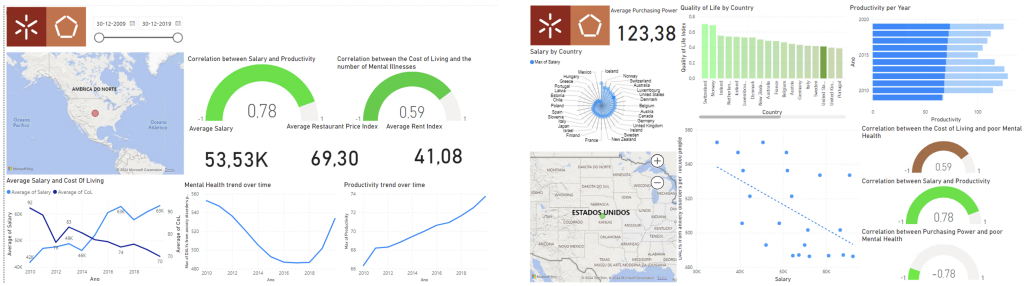

Fig. 7: United States case study

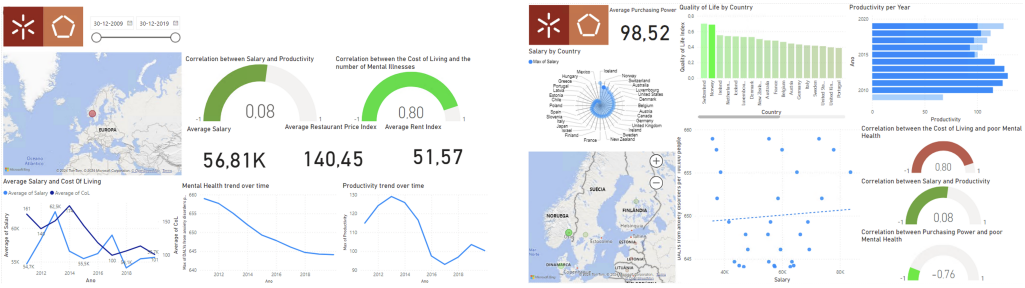

Fig. 8: Norway case study

From the dashboards above, we are able to get some important insights.

In the fig.7, we have the USA case study, overall the cost of living is strongly positively correlated with the number of mental illnesses in the USA. Throughout the years, in the USA, the salary and productivity have always been strongly correlated. We are also able to tell 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 USA population well being.

From the fig. 8, we can evaluate that in Norway from 2010 to 2019, indicating that as salaries rise, productivity tends to increase slightly. In contrast, the correlation between the cost of living and mental illness is much stronger, suggesting a nearly linear relationship where higher living costs are associated with more mental health issues.

Overall, we notice that globally, higher living costs are generally associated with an increase in mental health issues, while higher purchasing power tends to improve mental well-being. The relationship between salary and productivity varies, but in many cases, higher salaries can lead to increased productivity. These findings highlight the importance of considering economic factors such as cost of living and purchasing power when addressing mental health and productivity issues.

6 Conclusion

The study presented in this paper demonstrates a comprehensive approach to understanding the impact of purchasing power on mental health and productivity across different countries using big data analysis techniques. By meticulously selecting, preparing, and integrating various datasets, and employing robust data visualization tools, we have extracted valuable insights into the complex interplay between economic factors and socio-economic well-being.

6.1 Key Findings

Correlation Between Economic Factors and Mental Health:

- The analysis revealed significant correlations between purchasing power and mental health outcomes. Countries with higher purchasing power generally exhibited better

mental health indicators, suggesting that economic stability plays a crucial role in psychological well-being.

Productivity and Economic Stability:

- A direct relationship between higher productivity levels and stronger purchasing power was observed. This indicates that economic stability not only improves mental health but also boosts overall productivity, contributing to a nation's economic growth.

Importance of Comprehensive Data Integration:

- Integrating diverse datasets from different domains provided a holistic view, enabling a more accurate analysis of how economic factors influence mental health and productivity. The merged dataset approach proved efficient for querying and visualization, highlighting its efficacy for similar big data projects.

Efficiency of Data Transformation Tools:

- The benchmarks conducted between Pandas and PySpark for data transformation highlighted that, despite PySpark's capability for handling larger datasets, Pandas outperformed in terms of speed for our specific requirements. This emphasizes the importance of selecting the right tools based on the nature of the data and the task at hand.

6.2 Future Work

To build on our findings, future research could:

- Extend the Analysis Period: Including more recent data could provide updated insights into post-pandemic economic conditions and their effects on mental health and productivity.
- Incorporate Additional Variables: Exploring other socio-economic variables such as education levels, healthcare access, and social support systems could provide a more nuanced understanding of the factors influencing mental health and productivity.
- Develop Predictive Models: Utilizing machine learning techniques to predict trends in mental health and productivity based on economic indicators could offer proactive measures for policymakers.

Finally our study underscores the significant role of economic factors in shaping mental health and productivity outcomes. By leveraging big data systems and effective data management techniques, we have laid the groundwork for more informed and targeted socio-economic policies. The insights gained from this research can aid in the development of strategies aimed at improving both economic stability and public health.

References

[1] Agboola, A.A., Esan, O.T., Afolabi, O.T., Soyinka, T.A., Oluwaranti, A.O., Adetayo, A.: Economic burden of the therapeutic management of mental illnesses and

its effect on household purchasing power. PloS one **13**(9), 0202396 (2018)

[2] Oliveira, C., Saka, M., Bone, L., Jacobs, R.: 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)

[3] Bubonya, M., Cobb-Clark, D.A., Wooden, M.: Mental health and productivity at work: Does what you do matter? Labour economics **46**, 150–165 (2017)

[4] Numbeo: Cost of Living Rankings by Country. https://www.numbeo.com/cost-of-living/rankings_by_country.jsp (Accessed on 14 March 2024)

[5] Numbeo: Cost of Living - Consumer Price Index (CPI) Explained. https://www.numbeo.com/cost-of-living/cpi_explained.jsp. Accessed: March 14, 2024

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

[7] 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)

[8] Data, O.W.: Labor productivity per hour, Penn World Table. https://ourworldindata.org/grapher/labor-productivity-per-hour-pennworldtable?tab=table (Accessed on 14 March 2024)

[9] Zaman, A.U.: Pandas vs PySpark. https://medium.com/geekculture/pandas-vs-pyspark-fe110c266e5c. Accessed: March 25, 2024