Brazil Salaries Big Data Final Project

Problem Definition

This project explores salary data from Brazil's public sector employees to uncover patterns, detect inconsistencies, and derive key insights on income distribution and departmental spending. The dataset spans 2012 and 2013 and contains monthly, gross, and net salaries across different departments and job positions. Our goal is to:

- Identify salary distribution patterns across various positions and departments
- Understand deduction impacts (gross vs. net)
- Explore discrepancies between gross and net salaries.
- Addressed inconsistencies between roles and departments,
- Detect anomalies and duplicates
- Provide visualizations for deeper insights

Dataset Structure

Column Name	**Description**
name	Employee’s full name
department	Department the employee works in
month_total	Total monthly salary
gross_total	Salary before deductions
net_total	Salary after deductions
position	Job title or position
month	Month of the payment
year	Year of the payment

Tools & Technologies Used

- PySpark (data processing)
- Pandas (data wrangling)
- Matplotlib & Seaborn (visualizations)
- Google Colab
- GitHub (version control)

Methodology

1. Load & Explore
    - Load multiple CSVs from 2012 and 2013
    - Merge into one Spark DataFrame
    - Validate and cast data types
2. Data Cleaning
    - Ensured consistency by converting salary columns (month-total, gross_total and net_total) from string formats to numerical types.
3. Feature Engineering
    - Create new columns like DEDUCTIONS and DEDUCTION_PERCENTAGE
    - Assign numerical values to months for trend analysis
4. Aggregation & Analysis
    - Answer the predefined questions using grouping, filtering, and aggregation functions in Pandas.
    - Compute aggregates: average, total, max, min
5. Filtering & Ranking
    - Identify top earners, outliers, and salary brackets
    - Apply window functions like rank(), lag(), and lead()
6. Visualization
    - Create box plots, line graphs, and grouped bar charts

Analysis & Results

1. What is the distribution of salaries (net, gross, monthly)?

Most employees earn modest salaries between R$ 2,000 and R$ 4,000. A long tail of high earners skews the distribution.

2. Which departments or positions had the highest average earnings?

By grouping by department and position and calculating the average salary, we found that:

- **Top Departments:** Polícia Militar, Ministério Público
- **Top Positions:** Beneficiário de Seguridade, Delegado

3. Are there large disparities between gross and net salaries?

By comparing GROSS_TOTAL and NET_TOTAL, we noticed that in many cases, deductions were significant. The average deduction ranged between 20% to 40%, with some exceeding 50%.

4. Are there duplicate or suspiciously high salary records?

Yes. We noticed some suspicious values in the data like repeated names with same positions and net totals exist. Also filtered output of top 1% earners

5. How does salary vary across months from June to December?

We grouped salary totals by month and year to observe how payments changed over time. The salaries remained fairly consistent, with some noticeable increases near the end of the year, possibly due to bonuses or holiday pay.

6. What percentage of public funds went to the top 10% earners?

We identified the top 10% of employees by net salary and calculated how much of the total salary budget they received. This group received over 30% of total salary expenditure, emphasizing income inequality within the public sector.

7. Are there gender-based indicators in names?

We cautiously analyzed name endings like ‚-MARIA’ and ‚-JOAO’ but found this method too imprecise for reliable gender inference.

8. Can we create salary brackets and see distribution?

Yes. Employees were grouped by net salary brackets: <2000, 2000–4000, 4000–6000, etc. Most employees were within the 2000–4000 BRL range.

9. Do some departments have more employees, and how does that affect total spending?

Yes. For example, the Department of Education has many employees but lower average pay, while smaller departments like Judiciary spend more due to higher average salaries.

Challenges & Limitations

- Data inconsistencies (e.g., nulls, string-formatted numbers)
- No official gender data
- Department/job title standardization issues
- Didn’t perform a full join or self-join due to unclear matching keys
- PySpark used for loading, cleaning, aggregation. Pandas used for plotting.

Learnings

- Data cleaning is vital before any analysis
- Public sector salaries are highly skewed, with a small number of individuals earning a large share.
- Departments vary widely in terms of employee count and total spending.
- Net vs. gross discrepancies are significant and should be considered in salary evaluations.
- PySpark and Pandas can work hand-in-hand efficiently

Final Remarks

This analysis promotes transparency in public sector compensation and reveals how a small elite absorbs a large portion of salary spending. With cleaner data and better identifiers, future studies can go deeper and incorporate time trends or even prediction models.

Appendices

- GitHub Repo (Code + CSVs): https://github.com/EduardoEduBox/Salaries_Brazil_Analysis
- Notebooks: Eduardo.ipynb, Jolei.ipynb, Samantha.ipynb
- Slides (Canva): https://www.canva.com/design/DAGlrylDJs8/AAK1UAw630YQIg3skqA24A/edit?utm_content=DAGlrylDJs8&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton